HHS Public Access

Author manuscript

IEEE Int Conf Healthc Inform. Author manuscript; available in PMC 2020 June 13.

Published in final edited form as:

IEEE Int Conf Healthc Inform. 2019 June ; 2019: .

Detect Attributes of Medical Concepts via Sequence Labeling

Jun Xu,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Yang Xiang,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Zhiheng Li,
College of Computer Science and Technology, Dalian University of Technology, Dalian, China

Hee-Jin Lee,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Hua Xu*,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Qiang Wei,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Yaoyun Zhang,
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Yonghui Wu,
College of Medicine, University of Florida, Gainesville, Florida

Stephen Wu
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Abstract

In this study, we present a new method for detecting attributes of medical concepts, which uses a sequence labeling approach to recognize attribute entities and classify relations between concepts and attributes simultaneously within one step. A neural architecture combining bidirectional Long Short-Term Memory networks and Conditional Random fields (Bi-LSTMs-CRF) was adopted to

*indicates the corresponding author. Hua.Xu@uth.tmc.edu.

COMPETING INTERESTS

Dr. Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

detect disorder-modifier pairs in clinical text. Evaluations on the ShARe corpus show that the proposed method achieved higher accuracy and F1 scores than the traditional two-step approaches, indicating its potential to accelerate practical clinical NLP applications.

Keywords

information extraction; natural language processing; clinical notes

I. INTRODUCTION

Electronic Health Records (EHRs) are critical resources for clinical and translational research. Clinical Natural Language Processing (NLP) has been applied to extract and encode information in notes from EHRs. Although various clinical NLP approaches and systems[1], [2] have been developed to extract important medical entities from text, their associated attributes such as certainty, severity, etc. are also required by many downstream clinical applications. Recently several clinical NLP challenges focused on not only identifying medical concepts but also their associated attributes from clinical narratives. Many machine learning-based systems were developed in the challenges, and most of them used a two-step approach: 1) Named Entity Recognition (NER), to recognize attribute entities from text; and 2) Relation extraction, to classify the relations between any pair of attribute and target concept entities. Although these systems demonstrated reasonable performance on different attribute detection tasks, such approaches may suffer from error propagation, so that any errors generated in the NER step may propagate to the step of relation classification. In this study, we extended our previous work on drug-ADE detection in drug labels [3] to disease-attribute detection in clinical text, by developing a new deep learning-based sequence labeling approach (Bi-LSTM-CRF)[4], which recognizes attribute entities and classifies relations in one step.

II. METHODS

A. Task and Dataset

We used the ShARe corpus developed for the SemEval 2015 challenge task 14 [5], which is to recognize disorders and a set of attributes (Table I). For simplicity, we removed all disjoint disorder and attributes mentions and ignored the Generic indicator detection task since more than 99% of disorders have no Generic indicator attribute [5].

B. Traditional two-step approach (Baseline System)

The baseline system used a traditional two-step approach. In the first step we used a Bi-LSTM-CRF [4] to recognize attribute entities. Then a classifier was used to determine whether there's a relation between disorder and attribute entities. Two classifiers were used in the study: SVM and a deep neural network that combines a Bi-LSTM layer and a Softmax layer to classify candidate pairs [6].

C. Attribute detection by sequence labeling

Similar to our previous work on drug-ADR [3], we implemented a sequence labeling approach to recognize attributes and their relations to disorders here. We produced multiple training samples (named "concept-focused sequences" – CFS) from one sentence – one sample for each target concept (disorder). Figure 1 shows the transformed CFSs for target concepts "enlarged R kidney" and "air fluid level" respectively.

To model the target concept information alongside a CFS, we slightly modified the Bi-LSTM-CRF architecture by concatenating the vector representations of the target concept with the vector representations of individual words. We used "Target" and "NotTarget" tags to distinguish the target concept from other non-target concepts and embeddings of each tag was randomly initialized and learned directly from the data during the training of the model.

D. Experiments and Evaluation

We used the standard precision (P), recall (R) and F-measure as evaluation metrics. We also defined accuracy (Acc) defined as:

$$Acc = N_{correct\_predict} / N$$

Where, N is total number of gold standard concepts and $N_{correct\_predict}$ is number of concepts. For each task, we conducted 10-fold cross validation and reported micro-averages for each attribute type.

III. RESULTS AND DISCUSSION

The proposed sequence labeling approach greatly outperformed the traditional two-step approaches in both accuracy and F1 score (Table II), indicating the promise of this approach.

The disorder uncertainties (UNC) appears to be particularly difficult to detect, which may be due to diversity of the surface forms and low frequency of these attributes in our datasets. We also analyzed the errors of system prediction and identified several categories of errors: 1) The boundaries of the attribute entity did not perfectly match; 2) The system recognized an attribute entity and related it with a wrong target concept; 3) Annotation errors.

This study has several limitations. First, our Bi-LSTM-CRF system was not fully optimized for the problem setting. Second, our approach was only evaluated on a single corpus. In the future we will evaluate our approach on more corpora.

Acknowledgments

FUNDING

This study was supported in part by grants from NLM R01 LM010681, NCI U24 CA194215, and NCATS U01 TR002062.

REFERENCES

[1]. Xu J. UTH-CCB: The Participation of the SemEval 2015 Challenge-Task 14; Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); 2015. 311–314.

[2]. Wu Y, Jiang M, Xu J, Zhi D, and Xu H, "Clinical Named Entity Recognition Using Deep Learning Models.," in AMIA ... Annual Symposium proceedings. AMIA Symposium, 2017, vol. 2017, pp. 1812–1819.

[3]. Xu J, Lee H-J, Ji Z, Wang J, Wei Q, and Xu H, "UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017," in Proceedings of Text Analysis Conference.

[4]. Lample G, Ballesteros M, Subramanian S, Kawakami K, and Dyer C, "Neural Architectures for Named Entity Recognition," in Proceedings of NAACL-HLT, 2016, pp. 260–270.

[5]. Elhadad N. SemEval-2015 Task 14: Analysis of Clinical Text; Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); 2015. 303–310.

[6]. Zhang D and Wang D, "Relation classification via recurrent neural network," arXiv Prepr. arXiv1508.01006, 8 2015.

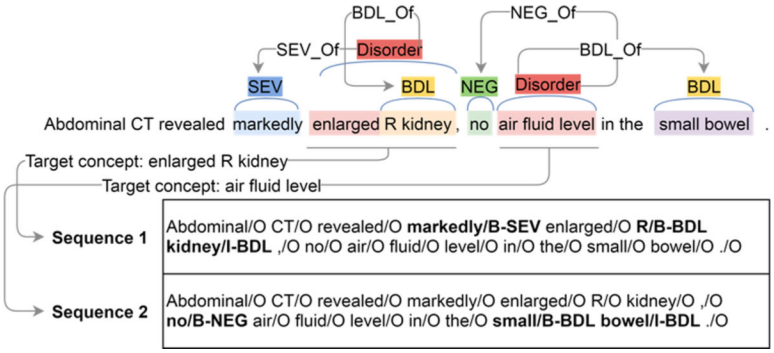

Fig. 1.

An illustration of the concept-focused sequence (CFS) transformation, where each separate sequence encodes all attributes for each target concept (Disorder).

TABLE I.

CONCEPTS AND ATTRIBUTES TYPES INCLUDED IN THIS STUDY, AS WELL AS THEIR DISTRIBUTION IN THE CORPORA.

# Disorder Concepts	#Attribute Mentions	
17,368	Negation indicator (NEG)	3,599
	Subject Class (SUB)	191
	Conditional indicator (CON)	927
	Severity class(SEV)	1,286
	Course class (COU)	901
	Uncertainty indicator (UNC)	1348
	Body location (BDL)	8,053

TABLE II.

THE OVERALL PERFORMANCE OF DIFFERENT APPROACHES ON THE SHARE-DISORDER DATASET IN DETECTING 7 ATTRIBUTES OF GIVEN DISORDERS: NEGATION (NEG), SUBJECT (SUB), CONDITIONAL (CON), SEVERITY (SEV), COURSE (COU), UNCERTAINTY (UNC), BODY LOCATION (BDL). BEST RESULTS ARE SHOWEN IN BOLDFACE.

Attribute		NEG	SUB	CON	SEV	COU	UNC	BDL
Baseline (Bi-LSTM-CRF+SVM)	Acc.	0.9323	0.9929	0.9669	0.9655	0.9576	0.9445	0.7524
	P	0.7931	0.7374	0.6990	0.6421	0.5068	0.4091	0.5887
	R	0.7768	0.6348	0.5987	0.7568	0.6437	0.4172	0.7516
	F	0.7849	0.6822	0.6449	0.6948	0.5671	0.4131	0.6602
Baseline (Bi-LSTM-CRF+Bi-LSTM)	Acc.	0.9146	0.9900	0.9632	0.9707	0.9597	0.9308	0.7859
	P	0.8387	0.8158	0.7872	0.7609	0.6340	0.4380	0.7218
	R	0.7277	0.5391	0.6054	0.8213	0.6322	0.3819	0.784
	F	0.7793	0.6492	0.6844	0.7900	0.6331	0.4080	0.7516
Sequence Labeling	Acc.	**0.9542**	**0.9937**	**0.9718**	**0.9817**	**0.9697**	**0.955**	**0.8695**
	P	0.8142	0.8222	0.7583	0.7812	0.6150	0.4854	0.7887
	R	0.8310	0.6435	0.6682	0.8859	0.7529	0.4393	0.7991
	F	**0.8225**	**0.7220**	**0.7104**	**0.8302**	**0.6770**	**0.4612**	**0.7939**